



인프라에서 Token Factory 로

AI 데이터센터 자체 구축 사업자의 핵심 격차와 해결 방안



목차

요약.....	2
1. 산업 배경: 데이터센터의 산출 단위가 바뀌고 있다.....	3
1.1 수용형 데이터센터에서 생산형 데이터센터로.....	3
1.2 동시에 진행되는 세 가지 구조 변화.....	3
1.2.1 추론이 학습을 넘어 주요 워크로드로 부상한다.....	3
1.2.2 과금 단위가 GPU 시간에서 토큰으로 이동한다.....	3
1.2.3 전력이 산출량의 상한을 결정한다.....	4
2. 자체 구축 운영의 네 가지 핵심 격차	4
2.1 컴퓨팅 자원 가동률 부족.....	5
2.2 멀티테넌시 및 거버넌스 역량 부족	6
2.3 사용량 측정 및 과금 체계의 부재	6
2.4 고객 온보딩 및 서비스 개시 소요 기간.....	7
소결	7
3. 운영 계층 참조 아키텍처와 AI-STACK / IXCSP 솔루션.....	8
3.1 운영 계층의 도입과 자체 구축에 대한 검토.....	8
3.2 3 계층 참조 아키텍처	8
3.3 AI-STACK과 IXCSP의 핵심 격차 대응 방안.....	9
3.4 자체 구축 사업자가 고려할 세 가지 관점	10
4. 도입 계획과 실제 사례.....	11
4.1 설비 생애주기에 맞춘 도입 단계.....	11
4.2 사례: 말레이시아 SAINS	12
5. 결론과 제언	12
상황별 권고.....	13

요약

인공지능의 확산은 데이터센터의 경제 구조를 바꾸고 있다. 추론 워크로드가 학습을 넘어 주요 워크로드로 자리 잡으면서, 데이터센터의 산출 단위도 랙과 전력에서 토큰으로 계량되는 서비스로 이동하고 있다. GTC 2026 에서 NVIDIA CEO 젠슨 황은 현대의 데이터센터를 산출 단위가 토큰인 공장으로 설명했다(Data Center Frontier, 2026). 딜로이트는 2026 년 추론이 전 세계 AI 컴퓨팅의 약 3 분의 2 를 차지할 것으로 전망한다(Deloitte, 2026).

이 변화는 AI 데이터센터를 직접 구축하는 사업자에게 새로운 문제를 던진다. 하드웨어와 설비 구축을 완료했다고 해서 곧바로 대외 서비스 제공 역량을 갖추는 것은 아니다. GPU 를 랙에 구축한 시점부터 첫 청구서를 발행하기까지는 이기종 컴퓨팅 자원 스케줄링, 멀티테넌시, API 연동, 사용량 측정·과금 정산이 필요하다. 그러나 이러한 운영 계층은 대부분의 구축 계획에 포함되지 않고 있다.

이러한 격차는 이미 업계 데이터에서도 확인되고 있다. Cast AI 가 약 23,000 개 기업용 쿠버네티스 클러스터의 실측 텔레메트리를 분석한 결과, 기업의 평균 GPU 가동률은 약 5%에 그쳤다(Cast AI, 2026). VentureBeat Research 가 기술 리더 573 명을 대상으로 실시한 조사에서는 자체 GPU 인프라를 운영하는 기업의 86%가 가동률이 절반 이하라고 답했다(VentureBeat Research, 2026). 또한 업계 분석에 따르면, 소비 계층(Consumption Layer)의 역량 격차로 인해 사업자는 통상 6~12 개월의 시간과 매출 기회를 잃을 수 있다(Supercomputing Frontiers, 2026).

본 백서는 AI 데이터센터를 계획하거나 구축 중인 사업자가 운영 계층의 핵심 격차를 식별할 수 있도록 돕고, 인피니틱스의 **AI-Stack** 과 **ixCSP** 가 컴퓨팅 자원을 매출로 전환하는 과정을 어떻게 단축하는지 보여 주기 위해 작성되었다. 본 백서는 네 부분으로 구성된다. 산업 배경, 네 가지 핵심 격차, 운영 계층 참조 아키텍처와 대응 솔루션, 그리고 도입 계획과 실제 사례이다.

1. 산업 배경: 데이터센터의 산출 단위가 바뀌고 있다

1.1 수용형 데이터센터에서 생산형 데이터센터로

지난 20년간 데이터센터의 핵심 역할은 스토리지와 컴퓨팅 자원을 수용·운영하는 것이었으며, 그 척도는 랙 수, 가용성, 전력사용효율(PUE)이었다. AI 워크로드의 부상은 이 정의를 바꿨다. 젠슨 황은 GTC 2026 미디어 브리핑에서 현대의 데이터센터를 산출 단위가 점차 토큰으로 전환되는 공장으로 설명하고, NVIDIA의 제품 로드맵을 토큰 처리량, 토큰 경제성, 와트당 성능과 연결했다. 이는 AI 인프라의 가치를 재는 새로운 잣대를 제시한 것으로, 핵심은 제한된 전력으로 얼마나 많은 실질적인 AI 가치를 창출할 수 있는가에 있다(Data Center Frontier, 2026). 이 틀에서 데이터센터는 전력과 데이터를 투입해 토큰을 산출하는 생산 시스템이며, 매출은 와트당 토큰 생산량과 가용 전력 용량의 곱으로 표현할 수 있다(spheron.network, 2026).

$$\text{매출} = \text{와트당 토큰 생산량} \times \text{가용 전력 용량(GW)}$$

1.2 동시에 진행되는 세 가지 구조 변화

1.2.1 추론이 학습을 넘어 주요 워크로드로 부상한다

딜로이트는 2026 TMT 전망에서 추론이 2026년 전 세계 AI 컴퓨팅의 약 3분의 2를 차지할 것으로 예측한다. 2025년 약 절반, 2023년 3분의 1에서 높아진 수치이며, 그 연산의 대부분은 여전히 엣지 단말이 아니라 데이터센터 내 고성능 AI 칩에서 수행된다(Deloitte, 2026).

데이터센터 관점에서 이는 워크로드의 특성이 변화한다는 의미다. 학습은 프로젝트 단위로 수행되며 매출이 일회성으로 발생하는 반면, 추론은 지속적인 워크로드로서 사용량에 따라 매출이 반복적으로 발생한다. 재무적으로 전자는 프로젝트성 매출, 후자는 반복 매출이며, 둘은 예측 가능성, 자금 조달 조건, 밸류에이션 논리에서 크게 다르다.

1.2.2 과금 단위가 GPU 시간에서 토큰으로 이동한다

NVIDIA의 기술 블로그는 Compute-as-a-service(CaaS) 계층이 시스템과 GPU 자원을 IaaS 형태로 시간 단위 제공하고, 그 위의 쿠버네티스 플랫폼 계층이 멀티테넌트 클러스터와 GPU 스케줄링을 제공하는 구조를 설명한다. 그러나 과금은 여전히 GPU 시간에 묶여 있다. 실질적인

비즈니스 모델의 전환은 사업자가 그 위에 토큰 사용량을 측정할 수 있는 모델과 애플리케이션을 구축하고, 인프라 사용 시간이 아닌 AI 서비스의 산출물을 판매하기 시작할 때 일어난다(NVIDIA Technical Blog, 2026).

이 모델은 이미 시장에 나와 있다. Omdia의 조사는 소버린 AI가 통신사업자의 AI 인프라 투자에서 가장 명확한 수익화 기회라고 지적하며, 중국 통신사들이 사용량 기반 AI 토큰 구독 요금제를 출시했고 이는 사업자가 AI 역량을 서비스로 구성하고 수익화하는 하나의 표준 모델이 될 수 있다고 본다(Omdia, 2026).

1.2.3 전력이 산출량의 상한을 결정한다

산출이 토큰으로 측정되고 용량이 가용 전력으로 제한될 때, 단위 전력당 생산할 수 있는 토큰의 양이 매출의 상한을 결정한다. 이 때문에 2026년에는 가동률이 단순한 엔지니어링 과제를 넘어 재무적 과제로 부상하고 있다. (NVIDIA Technical Blog, 2026).

2. 자체 구축 운영의 네 가지 핵심 격차

AI 데이터센터는 세 계층으로 나눌 수 있다. 설비·하드웨어 계층, 컴퓨팅 자원 관리 계층, 그리고 운영·수익화 계층이다. 자체 구축 사업자는 첫 계층에서 성숙한 역량을 갖추고 있고, 두 번째 계층은 제한적인 운영 역량으로 대응하고 있으며, 세 번째 계층은 아직 구축 계획에 포함되지 않은 경우가 많다.



그림 1. AI 데이터센터의 세 계층과 자체 구축 사업자의 역량 현황

NVIDIA 는 통신사업자용 AI 팩토리 레퍼런스 아키텍처를 설명하면서 인프라에서 토큰 계량형 서비스로 이행하는 데 필요한 역량을 모두 열거한다. 멀티테넌트 클러스터와 GPU 스케줄링, 모델 배포와 엔드포인트 관리, 구독하고 자동으로 프로비저닝할 수 있는 서비스 카탈로그, 그리고 토큰을 핵심 계량 지표로 활용하고 이를 성능·거버넌스·인프라 효율과 연결하는 계량·과금 계층이다. 이를 기반으로 할당량, 요청 속도 제한(Rate Limit), 서비스 수준 협약(SLA)을 적용한다(NVIDIA Technical Blog, 2026).

같은 목록은 이를 직접 구축하기가 얼마나 어려운지도 함께 보여 준다. 소버린 클라우드 플랫폼에 대한 업계 분석은 오픈소스 인프라 스택(OpenStack, Vanilla Kubernetes 등)이 완전한 통제권은 주지만 수익화 가능한 클라우드 운영 모델은 제공하지 않는다고 지적한다. 셀프서비스 포털, 사용량 측정, 통합 과금 기능을 갖춘 멀티테넌트 AI 플랫폼을 구축하려면 데이터센터를 짓거나 쿠버네티스 클러스터를 운영하는 것과는 상당히 다른 엔지니어링 전문성이 필요하며, 대부분의 사업자가 인프라 계층에서 멈추는 이유가 여기에 있다. 같은 분석은 소비 계층의 격차가 통상 6~12 개월의 시간과 매출 기회 손실로 이어지며, 인프라를 ‘보유한 상태’ 와 실제로 ‘서비스에 활용되는 상태’ 사이의 격차야말로 사업자가 시간과 매출 기회를 동시에 잃는 지점이라고 지적한다(Supercomputing Frontiers, 2026).

다음 네 가지 핵심 격차는 모두 두 번째와 세 번째 계층에서 발생한다.

2.1 컴퓨팅 자원 가동률 부족

Cast AI 의 2026 State of Kubernetes Optimization Report 는 약 23,000 개 기업용 쿠버네티스 클러스터의 프로덕션 텔레메트리를 분석해, 기업의 평균 GPU 가동률이 약 5%이고 프로비저닝된 용량이 실제 소비의 약 20 배임을 밝혔다(Cast AI, 2026). VentureBeat Research 가 2026 년 6 월 기술 리더 573 명을 대상으로 실시한 조사에서는 자체 GPU 인프라를 운영하는 기업의 86%가 가동률이 절반 이하라고 답했다(VentureBeat Research, 2026). 비교를 위해 덧붙이면, Cast AI 공동창업자 Laurent Gil 은 주야 주기와 통상적인 업무 리듬을 감안할 때 수동으로 운영되는 인프라의 현실적인 가동률 목표를 약 30% 수준으로 본다(VentureBeat, 2026). 주된 원인은 이기종 컴퓨팅 자원에 대한 통합 관리 체계의 부재, 가상화가 없어 단일 작업이 가속기 한 장을 통째로 점유하는 구조, 그리고 각 조직이 필요 이상으로 자원을 사전 할당받아 점유하는 관행이다.

영향: 투자 검토 단계에서 가정한 자원 가동률을 달성하지 못해 투자 회수 기간이 길어진다. 사용량 기반 가격 체계에서 유헤 인프라는 비용을 계속 발생시키며 유헤 산출 단위당 비용을 증가시킨다(VentureBeat, 2026).

2.2 멀티테넌시 및 거버넌스 역량 부족

고객이 단일 계정에서 다수로 확대되면 컴퓨팅·스토리지·네트워크 전반의 테넌트 격리와 함께 자원 할당량, 권한 통제, 운영 감사 로그, 컴플라이언스 보고가 필요해진다. 개별 기능 자체의 구현 난이도는 높지 않지만, 이 모두를 프로덕션 수준으로 끌어올리고 유지하는 일은 지속적인 엔지니어링 부담을 뜻한다. 소버린 AI와 정부 사업에서는 기준이 더 높다. Omdia는 데이터와 컴퓨팅에 대한 통제 수요 증가가 현지 사업자에게 유리한 환경을 만들고 있다고 지적하지만(Omdia, 2026), 데이터 레지던시(Data Residency), 접근 감사, 조직 간 자원 및 비용 배분은 있으면 좋은 요건이 아니라 필수 요건이다.

영향: 동시에 여러 고객을 수용하기 어려워 고객 집중 위험이 크게 남는다. 과금 체계도 다시 랙 단위 과금에 의존하게 되어, 실제 소비량에 따라 AI 서비스를 판매할 방법이 사라진다.

2.3 사용량 측정 및 과금 체계의 부재

토큰 사용량 측정은 GPU 시간 기반 측정과 요구되는 복잡도에서 본질적인 차이가 있다. 입력과 출력의 별도 가격 책정, 모델별로 다른 단가, 사용량 한도 관리 및 초과 사용 처리, 정산·대사 및 감사, 그리고 고객·재무·감사인을 위한 서로 다른 보고서 형식을 모두 다뤄야 한다. NVIDIA는 실질적인 경제적 전환이 인프라 위에 토큰 계량형 모델과 애플리케이션을 엮을 때 일어난다고 지적한다(NVIDIA Technical Blog, 2026). 통신 산업을 겨냥해 설계된 AI 토큰 과금 솔루션도 이미 등장해, 백오피스에서 측정된 사용량 데이터를 실제 판매 가능한 요금제로 전환하는 동시에 건별 실제 원가를 추적한다(Telecoms Tech News, 2026). 이 계층이 독자적인 제품 범주를 이룰 만큼 복잡하다는 방증이다.

영향: 네 가지 격차 중 매출에 가장 직접적인 영향을 미친다. 실무적으로는 월 정액제로 전환해 사용량 기반 과금의 가치를 포기하거나, 사용량을 수작업으로 산정하는 방법밖에 없다. 후자는 감사 근거가 없고 확장성이 떨어진다.

2.4 고객 온보딩 및 서비스 개시 소요 기간

물리적 구축 일정은 보통 프로젝트 관리 대상에 들어가지만, 설비 구축 완료 시점부터 실제 고객 서비스를 개시하기까지의 기간은 대부분의 계획에 포함되지 않는다. 이 기간에는 표준 API 및 API Gateway, 구독 및 자동 프로비저닝이 가능한 서비스 카탈로그, 고객 셀프서비스 포털, 모델 배포와 버전 관리, 개발자 문서가 들어간다. 이는 NVIDIA가 토큰 계량형 서비스에 필요하다고 열거한 역량과 일치한다([NVIDIA Technical Blog, 2026](#)).

이 기간의 재무적 의미는 하드웨어 감가상각이 검수일부터 시작되는 반면 매출은 운영 계층이 작동해야 비로소 발생한다는 데 있다. 설비 투자액이 3억 대만달러이고 5년에 걸쳐 감가상각한다고 가정하면 월 감가상각비는 약 500만 대만달러다. 운영 계층 가동이 6개월 늦어지면 — 앞서 인용한 6~12개월 구간의 하한이다 — 시장 기회비용을 빼고도 직접적인 감가상각 비용만 3,000만 대만달러에 이른다. 이 수치는 예시일 뿐이며, 실제 평가에서는 자사의 설비 투자액과 상각 기간을 대입해야 한다.

영향: Time-to-Revenue가 길어진다. 게다가 온보딩 경험에 대한 고객의 기대치는 퍼블릭 클라우드와 GPU 클라우드의 수분 내 셀프서비스 프로비저닝으로 이미 형성되어 있다.

소결

네 가지 핵심 격차의 공통점은 모두 데이터센터의 물리 인프라 영역이 아닌, 그 상위의 소프트웨어 계층에서 발생한다는 점이다.



그림 2. 네 가지 핵심 격차의 계층별 분포

3. 운영 계층 참조 아키텍처와 AI-Stack / ixCSP 솔루션

3.1 운영 계층의 도입과 자체 구축에 대한 검토

설비, 냉각 및 네트워크 시스템은 일반적으로 전문 업체의 솔루션을 도입하며, 운영 계층 역시 자체 구축보다 전문 솔루션 도입이 적합한 세 가지 특성을 가지고 있다. 첫째, 모델 세대와 고객 요구, 컴플라이언스 규정이 모두 계속 바뀌므로 장기적인 유지보수가 필요하다. 둘째, 특정 하드웨어 세대에 종속되지 않아 가속기 기종이 변경되더라도 플랫폼의 가치는 유지되며, 동일한 기능을 자체적으로 재개발하는 데 따른 추가적인 사업 가치는 제한적이다. 셋째, 규모에 따라 복잡도가 가파르게 올라가, 3 개 테넌트를 지원하는 시스템과 300 개 테넌트를 지원하는 시스템은 설계 복잡도에서 근본적인 차이가 있다.

3.2 3 계층 참조 아키텍처

자체 구축 또는 솔루션 도입 여부와 관계없이, Token Factory 를 운영하려면 세 계층이 필요하다. 아래 그림은 이를 피라미드로 제시한다. 아래에서부터 컴퓨팅 인프라 계층, AI 플랫폼 계층, AI 클라우드 운영 계층 순이며, 각 계층의 주요 책임을 오른쪽에 나열했다.



그림 3. Token Factory 운영 계층 참조 아키텍처

이 피라미드는 두 가지 특성을 반영한다. 첫째, 하위 계층은 상위 계층의 기반을 제공한다. 안정적인 자원 스케줄링 없이는 신뢰할 수 있는 모델 서비스를 제공하기 어렵고, 모델 서비스가 없으면 과금할 대상 자체가 없다. 둘째, 상위 계층으로 갈수록 해당 역량을 보유한 사업자는 줄어드는 반면, 매출 창출과의 연계성은 높아진다. 대부분의 자체 구축 사업자는 역량이 하위 인프라 계층에 집중되어 있으며, 이러한 역량 편중이 2 장에서

설명한 네 가지 핵심 격차가 발생하는 원인이다.

중요한 것은 세 계층이 모두 존재하는지만이 아니라 서로 연결되어 있는지다. 정확한 과금을 하려면 산출된 모든 토큰에 대해 어느 테넌트의 어느 모델이 어느 물리 자원으로 만들었는지 식별할 수 있어야 하며, 이러한 정보 연계 구조는 세 계층에 걸쳐 일관되게 유지되어야 한다. 연계가 어느 한 지점에서라도 단절되면 과금의 정확성과 감사 가능성을 확보하기 어렵다. 이는 여러 벤더의 제품을 조합할 때 가장 자주 부딪히는 어려움이기도 하다. 각 계층에 적합한 구성 요소를 확보하는 것보다, 이들을 하나의 완전한 정보 연계 구조로 통합하는 과정에 더 많은 엔지니어링 리소스가 필요하다. 비즈니스 모델 관점에서 컴퓨팅 인프라 계층만 갖춘 경우에는 코로케이션 중심의 인프라 제공에 머물 수 있으며, AI 플랫폼 계층까지 확장하면 GPU-as-a-Service 운영이 가능해진다. 여기에 AI 클라우드 운영 계층까지 연계되어야 Token Factory 로 확장할 수 있다.

3.3 AI-Stack 과 ixCSP 의 핵심 격차 대응 방안

ixCSP 는 인피니티스가 개발한 AI 클라우드 운영 플랫폼으로, 기업·통신사·데이터센터가 기존 GPU 서버 자원을 과금 및 거버넌스가 가능한 AI 클라우드 서비스 환경으로 전환할 수 있도록 지원한다. 이 플랫폼은 **AI-Stack 자원 관리 코어** 위에 구축되며, 그 위에 **AI Gateway** 와 **토큰 과금 시스템**을 통합한다. 3 계층 구조를 통해 GaaS, MaaS, TaaS 사업 모델을 지원하고, 사업자의 컴퓨팅 용량 계획에 따라 단계적으로 확장하거나 병행 운영할 수 있다.

AI-Stack 은 INFINITIX 가 오랜 기간 개발해 온 컴퓨팅 자원 관리 플랫폼으로, 소프트웨어 정의 GPU 와 스케줄링 기술로 가동률을 높이는 동시에 이기종 컴퓨팅 자원을 통합 관리한다. ixCSP 는 그 위에 멀티테넌시, 사용량 측정, 과금, 서비스 카탈로그 등의 클라우드 운영 기능을 제공한다. 이 역할 분담은 2 장의 격차 분포와 그대로 대응한다. 즉 **AI-Stack 은 컴퓨팅 자원 관리 계층의 자원 가동률 문제를, ixCSP 는 운영 계층의 서비스화 및 수익화 과제에 대응하며**, 두 제품은 프로젝트 진행에 맞춰 단계적으로 도입할 수 있다.

핵심 격차	대응 구성 요소 및 기능
2.1 컴퓨팅 자원 가동률	AI-Stack 자원 관리 코어: 소프트웨어 정의 GPU 와 스케줄링 기술로 GPU 가동률을 높이고, 이기종 컴퓨팅 자원을 통합 관리한다.
2.2 멀티테넌시와 거버넌스	ixCSP 멀티테넌시: 테넌트 격리, 자원 할당량, 권한 통제로 정부와 규제 산업의 거버넌스 요건을 지원한다
2.3 사용량 측정 및 과금	AI Gateway 는 표준 API 연동 지점이자 사용량 측정 지점으로 기능하며, 토큰 과금 시스템이 정확한 과금, 할당량 통제, 청구, 정산을 담당한다. 둘은 통합되어 과금 정보의 연계 구조를 일관되게 유지한다
2.4 온보딩과 서비스 개시	표준 API, 서비스 카탈로그, 모델 배포 역량으로 자체 개발에 필요한 엔지니어링 리소스를 줄이고 플랫폼 도입 기간을 단축한다.

3.4 자체 구축 사업자가 고려할 세 가지 관점

- **이기종 컴퓨팅 자원 지원:** 데이터센터는 운영 기간 동안 여러 세대의 가속기와 서로 다른 벤더의 제품을 함께 운영하게 된다. 특정 하드웨어에 종속된 플랫폼은 향후 조달의 유연성과 협상력을 제약한다. AI-Stack 의 이기종 관리는 구세대와 신세대를 나란히 운영할 수 있게 한다.
- **기존 자산 활용:** 기존 인프라를 활용해 도입할 수 있다. 유휴 상태이거나 가동률이 낮은 GPU 자원을 우선 통합 관리하여 운영 모델을 검증한 뒤, 용량 확장에 맞춰 규모를 키울 수 있다.
- **엔드투엔드 통합:** 이 플랫폼은 인프라 관리와 모델 배포부터 서비스 운영과 과금까지를 일관되게 포괄하므로 여러 솔루션을 개별적으로 연동·통합해야 하는 부담을 줄일 수 있다. 대규모 소프트웨어 엔지니어링 조직을 갖추지 못한 사업자에게 이는 도입 복잡성과 리스크를 줄이는 중요한 요소다.

4. 도입 계획과 실제 사례

4.1 설비 생애주기에 맞춘 도입 단계

데이터센터는 연 단위의 시간 축으로 구축된다. 따라서 도입 계획은 일반적인 소프트웨어 프로젝트의 단계 구분이 아니라 설비 생애주기에 맞춰 설계해야 한다.

단계	시점	주요 작업	흔히 놓치는 점
계획	투자 검토와 승인	사업 모델 조합을 정하고 가격을 설계하며, 운영 계층을 설비 투자에 포함한다	예산이 하드웨어와 설비에만 잡히고 운영 계층 항목이 없다
구축	설비 공사 기간	하드웨어 공사와 병행해 플랫폼을 배포하고 검증한다	소프트웨어를 맨 뒤로 미뤄, 설비는 완공됐지만 고객에게 서비스할 수 없다
가동 초기	초기 랙 가동	사내 워크로드와 핵심 고객을 대상으로 계량·과금 정확성을 검증한다	너무 이르게 멀티테넌트 규모를 쫓아 모든 문제가 한꺼번에 드러난다
확장	용량 확대	멀티테넌시를 개방하고 채널과 API 유통을 구축한다	용량은 늘렸는데 유통이 준비되지 않아 가동률이 떨어진다

계획 단계의 의사결정이 가장 큰 무게를 가지며, 운영 계층 예산은 이 시점에 확보해야 한다. 이유는 세 가지다.

- 예산 항목이 한번 확정되면 나중에 추가하기가 현저히 어려워진다
- 사업 모델 선택은 하드웨어 구성, 네트워크 및 스토리지 설계에 영향을 미친다
- 플랫폼 도입은 설비 공사와 병행할 수 있고, 그 병행 시간은 그대로 수익 실현까지의 시간 단축으로 이어진다.

가동 초기 단계의 실무적 권고는, 먼저 사내 워크로드를 대상으로 계량과 과금을 엔드투엔드로 검증해 계산이 맞고 보고서가 제대로 생성되는지 확인한 뒤 외부 고객에게 개방하는 것이다. 초기 과금 분쟁을 피하기 위해서다.

4.2 사례: 말레이시아 SAINS

인피니티스는 말레이시아 SAINS 와 양해각서를 체결하고 GPU 프라이빗 클라우드, AI 컴퓨팅 플랫폼, 기술 컨설팅, 토큰 이코노미 분야에서 협력하고 있다. 인피니티스는 AI-Stack 이기종 컴퓨팅 관리 플랫폼과 ixCSP AI 클라우드 운영 플랫폼을 제공해 정부와 기업이 GPU 프라이빗 클라우드와 AI 컴퓨팅 서비스를 구축하도록 지원한다. 두 플랫폼을 통합하면 AI 컴퓨팅을 AI 서비스와 토큰 가치로 전환하는 동시에 GPU 가동률을 높일 수 있다.

1991 년 설립된 SAINS 는 사라왁 주정부의 핵심 ICT 기술 파트너로, 1,100 명이 넘는 전문 인력을 보유하고 300 개 이상의 정부 정보 시스템을 구축해 왔다. 스마트시티, 디지털 트윈, 에이전트형 AI, 소버린 AI 인프라 사업을 추진하고 있으며, 최근에는 GPU-as-a-Service 와 LLM-as-a-Service 를 적극적으로 전개하고 있다. 이 사례는 전형적인 자체 구축 상황을 보여 준다. 규모도 기술력도 정책적 지원도 갖춰져 있고, 빠져 있는 것은 컴퓨팅을 서비스와 수익으로 바꾸는 운영 계층뿐이다.

5. 결론과 제언

데이터센터의 산출 단위는 랙에서 토큰으로 이동하고 있다. 2026 년 추론이 AI 컴퓨팅의 약 3 분의 2 를 차지할 것으로 전망되고, 과금 단위는 GPU 시간에서 토큰으로 이동하고, 전력 효율은 매출 잠재력을 결정하는 핵심 요소가 되고 있다. 이 흐름의 방향은 이미 분명하다.

그러나 하드웨어와 설비를 완성했다는 것이 곧 서비스 제공 역량을 뜻하지는 않는다. 기업의 평균 GPU 가동률이 약 5%이고 자체 구축 사업자의 86%가 가동률이 절반 이하라고 답한 현실이 보여 주는 것은 하드웨어의 부족이 아니라 운영 역량이 따라가지 못한다는 사실이다. 본 백서가 제시한 네 가지 격차는 모두 설비 위의 소프트웨어 계층에 있다.

그 계층은 장기적인 유지보수를 요구하고, 하드웨어 세대와 분리되어 있으며, 규모에 따라 복잡도가 가파르게 올라간다. 설비, 냉각, 네트워크 시스템과 같은 특성이며, 직접 만들기보다 성숙한 솔루션을 조달해야 하는 이유도 같다. AI-Stack 과 ixCSP 는 바로 이 두 계층을 위해 설계되었다. 전자는 기존 컴퓨팅 자원의 활용 효율을 높이고, 후자는 이를 과금 및 거버넌스가 가능한 AI 서비스로 전환한다.

상황별 권고

현재 상황	권고
투자 검토 단계	운영 계층을 아키텍처 계획에 명시적으로 포함하고, 그에 맞는 예산 항목이 있는지 확인한다. 이것이 없으면 재무 모델의 토큰 수익 가정에는 근거가 없다
구축 단계	플랫폼 평가와 도입은 설비 공사와 병행할 수 있다. 완공을 기다릴 필요가 없다
컴퓨팅은 갖췄으나 수익화하지 못한 경우	기존 자산을 활용한다. 먼저 AI-Stack 으로 가동률을 높이고, 이어서 ixCSP 를 도입해 계량과 과금을 검증한다
운영 중이며 고도화를 계획하는 경우	GaaS 에서 MaaS / TaaS 로 나아가는 데 필요한 것을 점검한다. 관건은 사용량 측정·과금 정보 연계 구조의 완결성이다